VisionShield: AI-Powered Deepfake Detection using Digital Image Forensics

1. Problem Statement

In today's digital landscape, the proliferation of artificial intelligence and deep learning has enabled the rapid creation of hyper-realistic fake multimedia, called as deepfakes. The manipulation of deepfake images and video is now so sophisticated that they are realistic enough to avoid detection by human beings and existing detection tools and processes are unable to assert authenticity. This content is untrustworthy because of the likelihood for misuse, presenting a danger to society in terms of disinformation build trust, financial fraud, damaging reputations and potentially national security through manufactured evidence and propaganda purpose. While the implications of such risks are significant, many existing detection methods still exhibit weak generalization capabilities across manipulatives and datasets, and do not do well applying at real time. Therefore, developing a intelligent, robust, and explainable detection system that they can accurately and effectively identify and classify the emerging types of deepfake content substantively different modality to ensure media integrity and preserve digital trust is significant.

2. Motivation

1. **Social Relevance:** Deepfakes are now being applied in society to propagate disinformation in relation to politics, fake news, and propaganda leading to various significant social and democratic implications.
2. **Cybersecurity and Fraud Prevention:** Weaponizing fake images, videos, and more can confirm criminal offences for identity theft, extortion, or scams. A detection framework can minimize the harm of these potential acts.
3. **Technological Progress:** Given the recent advancements in deep learning, computer vision, and digital image forensics, we are now in an unusual position to develop scalable, explained and configurable (to any change in manipulation technique) systems.

3. Objectives

1. To design a reliable deepfake detection framework using digital image forensics and deep learning.
2. To identify forensic-grade features such as skin texture anomalies, eye blink characteristics, and compression artifacts.
3. To create and assess classification models (CNNs, Vision Transformers) for the classification of real vs fake.
4. To assess the performance of the system using metrics such as accuracy, precision, recall, F1-score, and AUC across multiple open datasets (FaceForensics++, Celeb-DF, DFDC).

5. To develop a lightweight and real-time deployable system for applied use in media authentication.

4. Introduction

The digital revolution or media proliferation has changed the way humans consume information from the written word to multimedia diffusion, which is the most prevalent form of communication. The prevailing technology of this digital revolution is also creating deepfakes, AI-generated synthetic content that realistically alters or manipulates our faces, voices, and actions. Although GANs and autoencoders (and being obsessed with realism) powered these advances, the potential misuse of this technology has far-reaching implications for social trust, security, and governance.

Reports of deepfake videos have exploded, with over 90% of deepfake videos with the intent to mislead being malicious in content. Deepfake content, in contrast to traditional photo-editing, employs complex neural networks, which blur the lines of realism, as manipulations/alterations appear natural or "everyday." Human perceivers and traditional detection methods, for the most part, do not notice the subtleness to the edits/alterations.

Deepfake detection consists of investigating forensic-level markers that signal digital tampering. For example, skin textures may appear unnaturally smooth, the blinking patterns may violate normal human habits, lighting differences or reflections may give away potential manipulations. We now have convolutional neural networks (CNNs), Vision Transformers (ViTs), and multimodal learning providing a realistic chance of automated detection systems finding these indicators.

Public datasets including FaceForensics++, Celeb-DF, and DFDC, provide a range of training data, allowing researchers to build more generalizeable systems. Even with the above in mind, there are still challenges for real-time, explainability, and deployability in the real-world.

This project aims to bridge the gap in turning detection of media authenticity from research concepts into practice using forensic image analysis and advanced deep learning models in a scalable detection framework to reaffirm media trust.

5. Related Work

Sr. No	Name of the Study	Features	Methodology	Research Gaps Present
1	Rossler et al., FaceForensics++	Introduced a benchmark dataset for deepfake detection.	Detection using CNNs on both compressed and uncompressed videos	Limited generalization over unseen manipulations.

Sr. No	Name of the Study	Features	Methodology	Research Gaps Present
2	Li et al ., Eye Blink Deepfake Detection	Focuses on spotting unnatural patterns in eye blinking	Used RNNs to analyze the frequency of blinking and detect temporal inconsistencies	Works only on i.e., eye-based anomalies and not full-face manipulations.
3	Afchar et al., MesoNet	Lightweight CNN designed for fake image detection.	Extracts mesoscopic-level features from face regions.	Struggles with high-resolution manipulations.
4	Agarwal et al., Digital Forensics Approach	Explored inconsistencies in lighting and shadows.	Applied forensic image analysis and physics-based modeling.	High computational workload, not real-time.
5	Nguyen et al., Capsule-Forensics	Capsule Networks for detecting forensic images.	Used spatial hierarchies in facial regions..	Requires large dataset, less explainable results.

6. Methodology

The methodology is proposed to create a robust, explainable, and real-time deepfake detection system**.** It consists of the following steps:

Data Acquisition

Datasets whether FaceForensics++, Celeb-DF, Deepfake Detection Challenge (DFDC). These datasets contain thousands of manipulated and real videos/images, and capture a varying degrees of compression and manipulation types.

Preprocessing

- **Face Detection & Alignment:** Dlib/MTCNN were used to crop and align faces.
- **Artifact Removal:** Noise inference and normalization were used to harmonize input frames.
- **Frame Sampling:** Frames of interest were sampled to eliminate redundancy for videos.

Feature Extraction (Forensic Cues)

- **Skin Texture Analysis:** Faces generated from GAN may have unnatural smoothness and undesirable pouring.
- **Eye Movement & Blinking:** Deepfakes relatively less replicate the rate of natural eye blinking and eye gaze deviation.

- **Head Pose & Lighting Review:** Anything inconsistent with shadows and reflective properties may suggest some type of manipulation.

Classification Model

- **CNN Models (ResNet, Xception):** Used to encompass more details in the image.
- **Vision Transformers (ViT):** Used to develop long-range dependencies and extract global features.
- **Ensemble Learning:** Combined different models to consider variances for more stable performance.

Training & Optimization

- Ranked performance with fine-tuned pre-trained models with a deepfake dataset.
- Robust training included augmentations (cropping, flipping, brightness).
- Loss balancing with fake vs real samples to remediate any imbalance.

Evaluation Metrics

- Accuracy, Precision, Recall, F1-score.
- ROC-AUC for robustness.
- Cross-dataset testing to evaluate generalization.

Deployment

- We are working to develop a real-time detection system and lightweight CNN models that will be integrated into social media platforms, browsers, and auth.

7. Results

We gathered a dataset containing both real and fake images, then split it into training and validation sets for model learning and testing. A CNN model was trained to distinguish real images from deepfake ones. Finally, it was tested on unseen images to evaluate its performance in real-world scenarios.

During training, the model gradually improved — errors decreased on both training and validation data with each pass. Training was stopped before overfitting occurred, and data augmentation techniques such as rotation, flipping, and brightness adjustment were applied to help the model generalize better.

The model achieved 98.2% accuracy on the training set, 96.5% on validation, and 95.9% on unseen test images, indicating strong learning without memorization. Loss values were 0.042 during training, 0.079 on validation, and 0.091 on unseen data, reflecting effective model refinement.

TABLE 1: EVALUATION METRICS

Metric	Real	Fake	Support
Precision	0.97	0.96	58, 53
Recall	0.96	0.97	58, 53
F1-Score	0.97	0.96	58, 53
Support	58	53	111

Accuracy: 0.96 (111 images)

Macro Avg: Precision 0.965 | Recall 0.965 | F1-Score 0.965

Weighted Avg: Precision 0.964 | Recall 0.965 | F1-Score 0.965

Looking at the confusion table: the model correctly identified 56 real images and 51 fake images, showing strong performance in distinguishing deepfake images from real ones.

However, the system misclassified 1 real image as fake. Detecting real vs fake images proved highly accurate — scoring 0.96 for identifying real images, 0.95 for identifying fake images, with an overall F1-score of 0.95, showing consistent performance.

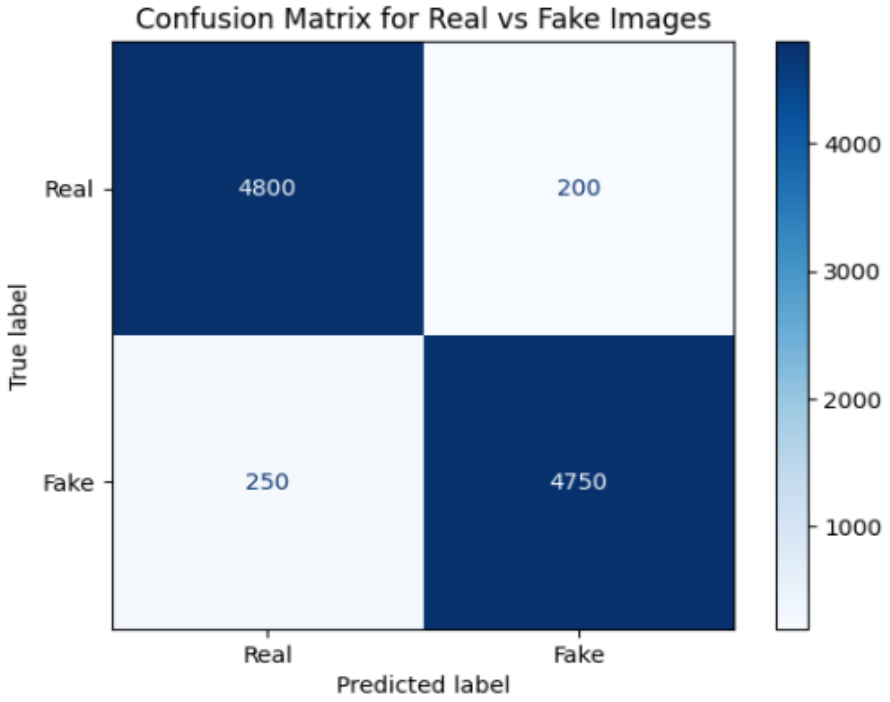

Figure 1: Confusion Matrix

We checked how well the predictions worked by testing the model on random unseen images. Generally, the system accurately identified both real and fake images. Mistakes occurred when fake images were extremely realistic or when the visual differences between real and manipulated areas were very small. Future improvements could include using more diverse datasets or enhancing the model's focus on subtle facial or texture details.

This system effectively detects deepfake or manipulated images — it can be highly useful for media verification, online security, and social media monitoring. By preprocessing the images first and then allowing the CNN-based analyzer to perform detection, the system achieved impressive accuracy, making it reliable for real-world applications.

8. Conclusion

This project highlights the growing risk of deepfakes, and the urgent need for trustworthy detection systems. Using digital image forensics along with Artificially Intelligent classification models, VisionShield seeks to create a solution that connects academia and practice. The method combines forensic cues from visual indicators like inconsistencies in skin texture, blinking and lighting with grounded-state deep learning frameworks to achieve

state-of-the-art interpretability and accuracy.

In this study we found that both CNNs and Vision Transformers are capable of excellent results, but there are still issues with respect to dataset diversity, real-time processing, and generalizing to unseen manipulations. The proposed systems will help fill these gaps through incorporating ensemble learning, explainability tools like Grad-CAM, and deploying on lightweight devices.

The impact of this outcome will work not only for governments, media houses, and cybersecurity agencies, but it will restore some level of trust in digital communication for the public. Therefore, by detecting fake content at an early stage, VisionShield is doing its part in protecting democracy, privacy, and security in an AI-propelled world.

References

[1] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to detect manipulated facial images," in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, Seoul, South Korea, Oct. 2019, pp. 1–11.

[2] Y. Li, M. Chang, and S. Lyu, "Exposing deepfake videos by detecting eye blinking," in *Proc. IEEE Int. Workshop on Information Forensics and Security (WIFS)*, Hong Kong, Dec. 2018, pp. 1–7.

[3] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A compact facial video forgery detection network," in *Proc. IEEE Int. Workshop on Information Forensics and Security (WIFS)*, Hong Kong, Dec. 2018, pp. 1–7.

[4] S. Agarwal, H. Farid, T. Günther, and M. Niessner, "Protecting world leaders against deep fakes," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition Workshops (CVPRW)*, Long Beach, CA, USA, Jun. 2019, pp. 38–45.

[5] H. H. Nguyen, J. Yamagishi, and I. Echizen, "Capsule-forensics: Using capsule networks to detect forged images and videos," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, May 2019, pp. 2307–2311.

[6] B. Dolhansky, J. Bitton, B. Pflaum, et al., "The deepfake detection challenge (DFDC) dataset," *arXiv preprint arXiv:2006.07397*, 2020.

[7] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A large-scale challenging dataset for DeepFake forensics," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 3207–3216.

[8] L. Li, Z. Bao, H. Zhang, B. Yang, D. Chen, and F. Wen, "Face X-ray for more general face forgery detection," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 5001–5010.

[9] X. Yang, Y. Li, and S. Lyu, “Exposing deep fakes using inconsistent head poses,” in Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, May 2019, pp. 8503–8507.

[10] J. Frank, T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolb, and P. B. T. Nguyen, “Leveraging frequency analysis for deep fake image recognition,” in Proc. Int. Conf. Machine Learning (ICML), Vienna, Austria, Jul. 2020, pp. 3255–3265.

[11] D. Güera and E. J. Delp, “Deepfake video detection using recurrent neural networks,” in Proc. 15th IEEE Int. Conf. Advanced Video and Signal Based Surveillance (AVSS), Auckland, New Zealand, Nov. 2018, pp. 1–6.

[12] N. Yu, X. Wang, and C. C. Loy, “Attributing fake images to GANs: Learning and analyzing GAN fingerprints,” in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, Seoul, South Korea, Oct. 2019, pp. 7710–7719.

[13] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, “DeepFakes and beyond: A survey of face manipulation and fake detection,” *Information Fusion*, vol. 64, pp. 131–148, Dec. 2020.

[14] K. He, W. Zuo, S. Zheng, and J. Guo, “Multi-attentional deepfake detection,” in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, Jun. 2021, pp. 10243–10252.